A nearly-optimal method to compute the truncated theta function, its derivatives, and integrals

Ghaith Ayesh Hiary

February 2, 2008

Abstract

A polynomial-time method to compute the truncated theta function, its derivatives, and integrals is presented. The method is elementary, rigorous, explicit, and suited for computer implementation. The basic idea is to iteratively apply Poisson summation while suitably normalizing the arguments of the truncated theta function in each iteration. The method relies on the periodicity of the complex exponential, which enables the normalization of the arguments, and the self-similarity of the Gaussian, which allows for repeated application of Poisson summation; in other words, the method relies on modular properties of the theta function. Applications to the numerical computation of the Riemann zeta function and to finding the number of solutions of Waring type Diophantine equations are presented.

1 Introduction

Sums of the form

$$\sum_{k=K_1}^{K_2} g(k)\exp(f(k)), \qquad f(x)\in\mathbb{C}[x] \tag{1}$$

arise in areas such as number theory, differential equations, lattice-point problems, optics, and mathematical physics, among others. For example, one encounters such sums in the context of Diophantine equations and fractional parts of polynomials ([Ko]), solutions of heat and wave equations ([M]), counting of integer points lying close to a curve ([Hu]), numerical integration and quadrature formulas ([Ko]), and motion of harmonic oscillators ([Ka]). Due to the importance such sums, there exists an abundance of methods to bound them. For instance, Vinogradov's [V] methods supply bounds to these sums, which, along with some involved sieving techniques, are used in attacking Goldbach-Waring type problems (see [LWY] for example).

Despite the substantial interest in such sums, comparatively little is known about how to compute them for general values of their arguments. Yet, in some situations, it is useful to be able to compute such sums efficiently. We later describe two such settings, both of which originate in number theory.

The simplest examples of (1) occur when $f(x)$ is of degree one and $g(x)$ is a polynomial. There, we obtain the geometric series and its derivatives, where "closed-form" formulae for their values are available. The first non-trivial example occurs when $f(x)$ is a quadratic and $g(k)$ is a polynomial. Let us specialize to this case. Then the sums to be evaluated can be reduced to the form

$$F(K, j; a, b) := \frac{1}{K^j} \sum_{k=0}^{K} k^j \exp(2\pi i a k + 2\pi i b k^2) \tag{2}$$

In this paper, using ideas that are rooted in analysis, we prove that $F(K, j; a, b)$ can be numerically computed in logarithmic time in K for any a, b, and $j \geq 0$. For brevity, we sometimes refer to this result as the "theta algorithm." Let us demonstrate the utility of our algorithm.

For many reasons, which include numerically verifying the Riemann Hypothesis, moment questions, and, more recently, connections to Random Matrix Theory models, the values of $\zeta(1/2+it)$ on finite intervals are of great interest to number theorists. There exist several methods to *compute* $\zeta(1/2+it)$ *with polynomial accuracy*; that is, methods to obtain the numerical values of $\zeta(1/2+it)$ for $t > t_0$, t_0 some fixed positive number, with absolute error bounded by t^{-c} for fixed $c > 0$. An elementary method is usually derived from the Euler-Maclaurin summation formula; see [E]. It has complexity $O(t)$. Another method relies on the straightforward application of the Riemann-Siegel formula. One frequently used version of the formula is

$$\zeta(1/2+it) = \Re\left(2e^{i\theta(t)} \sum_{n=1}^{n_1} n^{-1/2} \exp(it \log n)\right) + \Phi(t) + O(t^{-5/4}) \tag{3}$$

where $n_1 := \lfloor \sqrt{t/(2\pi)} \rfloor$, and $\theta(t)$, $\Phi(t)$ are certain real-valued functions that can be evaluated in time $O(\log t)$; see [E].

More recently, Odlyzko and Schönhage [OS] derived a practical algorithm to *simultaneously* compute $O(T^{1/2})$ values of $\zeta(1/2+iT+it)$, where $t \in [0, T^{1/2}]$, in $T^{1/2+o(1)}$ time. The Odlyzko-Schönhage algorithm did not reduce the cost of a *single* evaluation of zeta. A single evaluation still consumed $O(t^{1/2})$ time using the Riemann-Siegel method. Then, Schönhage [S] lowered the cost of a single evaluation of zeta to $t^{3/8+o(1)}$. Later, Heath-Brown [HB] presented ideas that further improved the complexity of a single evaluation to $t^{1/3+o(1)}$.

Our theta algorithm directly leads to another and potentially practical method to compute $\zeta(1/2+it)$ at a single point with polynomial accuracy in $t^{1/3+o(1)}$ time. The derivation is explained in a general context in [H] and [S] (similar manipulations can also be found in [T], page 99). The basic idea is to apply appropriate subdivisions and Taylor expansions to the main sum in the

Riemann-Siegel formula in order to reduce its computation to that of a sum of $O(t^{1/3})$ terms of the form $F(K, j; a, b)$.

As another simple and direct application of the theta algorithm, we show how to find the number of solutions of a Waring type Diophantine equation. Suppose we want to find the number of integer solutions to the system

$$\sum_{i=1}^{s}(\alpha_i k_i + \beta_i k_i^2) - \sum_{i=s+1}^{s+t}(\alpha_i k_i + \beta_i k_i^2) \equiv 0 \,(\mathrm{mod}\ M) \tag{4}$$
$$0 \le k_1, \ldots, k_{s+t} \le K, \qquad \alpha_1, \beta_1, \ldots, \alpha_{s+t}, \beta_{s+t} \text{ are some fixed integers}$$

A simple calculation reveals that the number of solutions is given by

$$\frac{1}{M}\sum_{l=0}^{M-1}\left(\prod_{i=1}^{s} F(K, 0; \alpha_i\, l/M, \beta_i\, l/M)\right)\left(\prod_{i=s+1}^{s+t}\overline{F(K, 0; \alpha_i\, l/M, \beta_i\, l/M)^t}\right) \tag{5}$$

Since we can compute F in polynomial time, the cost of computing (5) is $(s+t)MK^{o(1)}$. This is already significantly better than a brute-force method. One can use the Fast Fourier Transform to compute (5) with sufficient accuracy in $(s+t)MK^{o(1)} + K^{3+o(1)}$ time. But this is less efficient and it requires temporarily storing large amounts of data. In the special case $M = K$, one can use Gauss sums to solve the problem in complexity $O((s+t)M)$.

To gain a better appreciation of the "generic" behavior of $F(K, j; a, b)$ it is useful to consider some numerical evidence. Figure 1 depicts the real part of $F(100, 0; a, b)$ where either a or b runs over the interval $[.110, .113]$.

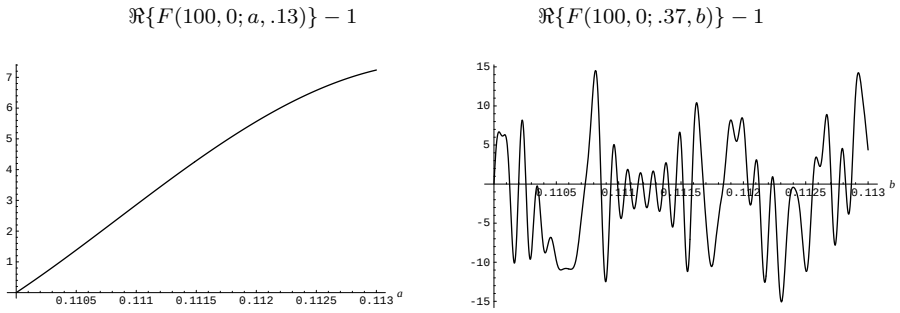

Figure 1: Two plots of $\Re\{F(100, 0; a, b)\} - 1$ for selected values.

Since the argument of $e^{2\pi iak + 2\pi ibk^2}$ is more sensitive to perturbations in b than a, it is not surprising that F exhibits much less uniformity as b varies. Also, the size of $\Re\{F(100, 0; .37, b)\} - 1$ is on the order of $\sqrt{200} \approx 14.1$. This is consistent with the behavior of a sequence of independent random variables of mean zero and variance $1/2$. Note however that around rationals with relatively low denominators the behavior is markedly different; there, sudden spikes or troughs may occur (e.g. $F(K, 0; 1/2, 1/2) = K + 1$).

In searching for methods to compute $F(K, j; a, b)$, one should capitalize on the rich structure of the theta function. The theta function, together with variants, occurs frequently in number theory; it is directly related to the zeta function by a Mellin Transform, it has a functional equation, and other modular properties. Thus, one anticipates that a promising method to compute a truncated theta function would be one that directly integrates such features in its design. Indeed, this viewpoint leads to a natural solution of the problem. Before we embark on a detailed description of the solution, we state it as

Theorem 1.1. *There exist absolute constants κ_1 and κ_2 such that for any $\epsilon \in (0, e^{-1})$, any positive integer K, any non-negative integer j, any $a, b \in \mathbb{R}/\mathbb{Z}$, and with an underlying computational model that performs arithmetics using $O(\nu^2)$ bits, where $\nu = (j+1)\log(K/\epsilon)$, the function $F(K, j; a, b)$ can be computed with error bounded by $O(\nu^{\kappa_1}\epsilon)$ using $O(\nu^{\kappa_2})$ arithmetic operations. The big-O constants are absolute.*

See the beginning of Section 3 for the definition of an *arithmetic operation*. Also, note that a bit complexity bound follows routinely from the arithmetic operations bound because all the numbers that occur in our algorithm have $O(\nu^2)$ bits. We also show how to compute the related sums

$$G(K, j; a, b) := \sum_{k=1}^{K} \frac{1}{k^j} \exp(2\pi i a k + 2\pi i b k^2)$$

with similar error and complexity. This is done mainly for use in the separate paper [H].

We do not try to obtain numerical values for the constants κ_1 and κ_2 in Theorem 1.1. With some optimization, they could probably be taken to be around 3. Also, in a practical version of the algorithm the arithmetic can be performed using substantially fewer than $O(\nu^2)$ bits and we would likely be able to replace $\nu(K, j, \epsilon)$ with $j + \log(K/\epsilon)$.

Let us motivate our method to compute $F(K, j; a, b)$ when $j = 0$. To this end, recall the following application of Poisson summation due to van der Corput; see [T], page 74, for a slightly different but equivalent version. We refer to this application as the *van der Corput iteration*, although it is not conventionally labelled as such.

Theorem 1.2 (van der Corput iteration)**.** *Let $f(x)$ be a continuous real function with monotonically increasing derivative in (s, t). Let $f'(s) = \alpha$, $f'(t) = \beta$. Then*

$$\sum_{s<k\le t} \exp(2\pi i f(k)) = \sum_{\alpha-\eta<v<\beta+\eta} \int_s^t \exp(2\pi i (f(x) - vx))\, dx + R_1 \qquad (6)$$

where $0 < \eta < 1$ and $R_1 = O(\log(\beta - \alpha + 2))$.

The van der Corput iteration converts a sum of $t - s$ terms to a sum of $\beta - \alpha$ terms. In order to turn this transformation into a potentially useful computational device, we almost surely need $\beta - \alpha \le c(t - s)$ for some absolute constant

$0 \leq c < 1$. Moreover, we must be able to compute R_1 and each of the terms in the sum over v, which are precisely the terms in the Poisson summation formula that contain critical points, using relatively few operations. Still, if $c > 0$, the length of the sum over v may be of the same order of magnitude as the length of the original sum. Thus, the complexity of the problem appears unchanged unless we can handle the sum over v efficiently. However, if we require the function $e^{2\pi i f(x)}$ to possess some favorable Fourier transform properties that allow us to turn the v-terms into ones suited for yet another application of (6), then, under such hypotheses, one may hope that repeated applications of the van der Corput iteration are possible. If they are, one can compute the original sum over k in logarithmic time in K.

Such restrictions on $f(x)$ are quite stringent, which will severely limit potential candidates for the proposed strategy. Fortunately, the choice $f(x) = ax + bx^2$ is particularly amenable to repeated applications of the van der Corput iteration. Indeed, let $s = 0$ and $t = K$. Assume that $\lceil a \rceil \leq \lfloor a + 2bK \rfloor$, which is frequently the case. Then, the relation (6) becomes

$$\sum_{k=0}^{K} \exp(2\pi iak + 2\pi ibk^2) = \sum_{v=\lceil a \rceil}^{\lfloor a+2bK \rfloor} \int_0^K \exp(2\pi iax + 2\pi ibx^2 - 2\pi ivx)\, dx + R_1 \quad (7)$$

Let us write (7) as $S_l = S_r + R_1$. We refer to sums of the form S_l as "quadratic" sums. First, recall the following "self-similarity" property of the Gaussian;

$$\int_{-\infty}^{\infty} e^{\alpha t - t^2}\, dt = \sqrt{\pi} e^{\alpha^2/4}, \qquad \alpha \in \mathbb{C}$$

The method that we develop works as follows. It is easily shown that we can always reduce to $a \in [0, 1)$ and $b \in [0, 1/4]$ in (7). This is important, otherwise consecutive applications of Poisson summation essentially cancel each other out. But with $b \in [0, 1/4]$, S_r has length $\leq K/2$. Furthermore, using the self-similarity of the Gaussian as well as shifting of integrals to their stationary phase, the sum S_r is transformed to a quadratic sum of the same length plus a remainder R_2. Finally, R_1 and R_2 are computed quickly. By iterating this procedure at most $\log_2 K$ times, we arrive at a sum of length $O(1)$ that can be evaluated directly. In other words, most of the work of the algorithm consists of computing the aggregate of the "errors" R_1 and R_2. Note that in this framework, varying a corresponds to sliding the sum over v whereas varying b corresponds to compressing, or stretching, the sum. The latter feature largely accounts for applicability of the van der Corput iteration in the context of this paper.

For general $j \geq 0$, the method consists of at most $\log_2 K$ iterations. Each iteration acts on $F(K, j; a, b)$ in the following way

$$F(K, j; a, b) = \sum_{l=0}^{j} w_{l,j,a,b} F\left(K^*_{a,b,K}, l; a^*_{a,b}, b^*_{a,b}\right) + R_{K,j,a,b} + E_{K,j,a,b} \qquad (8)$$

Let us suppress dependencies to avoid notational clutter. Then, it is shown that there exist absolute constants $\tilde{\kappa}_1$ and $\tilde{\kappa}_2$ such that R and the coefficients $w_{l,j}$ can be computed using $O(\nu^{\tilde{\kappa}_2})$ operations, the function $E_{K,j,a,b} = O(\nu^{\tilde{\kappa}_1}\epsilon)$, and $K_1 \leq K/2$. A key point is that the tuple (K^*, a^*, b^*) does not depend on j. Therefore, the number of sums F we need to compute in each iteration is always $\leq j+1$. To elaborate, our method acts on a sum of the form $\sum_{l=0}^{j} z_{l,j} F(K, l; a, b)$ in the following way

$$\sum_{l=0}^{j} z_l F(K, l; a, b) = \sum_{l=0}^{j} \tilde{w}_{l,j,a,b} F\left(K^*_{a,b,K}, l; a^*_{a,b}, b^*_{a,b}\right) + \sum_{l=0}^{j} R_{K,l,a,b} + \sum_{l=0}^{j} E_{K,l,a,b}$$

where

$$\tilde{w}_{l,j,a,b} := \sum_{s=l}^{j} z_s w_{l,s,a,b} \tag{9}$$

and $w_{l,j,a,b}$ are defined as in (8). We also show that the coefficients $w_{l,j}$ do not grow too rapidly with each iteration in the algorithm. In fact, in Section 3 we show that the maximum modulus of $w_{l,j}$ over all iterations of the algorithm is $O((j+1)4^j K^2)$. This bound is rather generous but it is sharp enough for purposes of our error analysis. As for the sums $G(K, j; a, b)$, they will be reconstructed as short linear combinations of sums of the form F.

The paper is organized as follows. For easy reference, Section 2 lists various contours, integrals, and other quantities defined in the paper. Section 3 describes the typical van der Corput iteration. Each such iteration amounts to evaluating certain "short" exponential integrals, which in turn are converted to sums. Section 4 gives pseudo-code for the algorithm. Section 5 shows how to compute the related sums $G(K, j; a, b)$. Finally, Section 6 contains proofs of various lemmas employed in previous sections. We also decided to include lemmas 6.7 and 6.8 in Section 6. Although these lemmas are primarily intended for use in the separate paper [H], it seems that their natural context is here as they are merely specializations of the theta algorithm.

2 Notation

For easy reference, we list expressions defined in this paper. Define the contours

$$\begin{aligned} C_0 &:= \{t \,|\, 0 \leq t \leq K\}, & C_1 &:= \{K + it \,|\, 0 \leq t \leq K\} \\ C_2 &:= \{te^{\pi i/4} \,|\, 0 \leq t \leq \sqrt{2}K\}, & C_3 &= \{-it | 0 \leq t < \infty\} \\ C_4 &= \{K - it | 0 \leq t < \infty\}, & C_5 &:= \{te^{\pi i/4} | -\infty < t < 0\} \\ C_6 &:= \{te^{\pi i/4} | \sqrt{2}K < t < \infty\}, & C_7 &:= \{te^{-\pi i/4} \,|\, 0 \leq t \leq \sqrt{2}K\} \end{aligned}$$

Also, let $C_8 := C_2 \cup C_5 \cup C_6$ and $C_9 := \{t \,|\, 0 \leq t \leq \infty\}$. Figure 2 depicts these contours. Note that dots and dashed lines determine boundaries of contours; the latter represents lines that continue indefinitely; we make use of the usual set difference operator "\" when labelling contours.

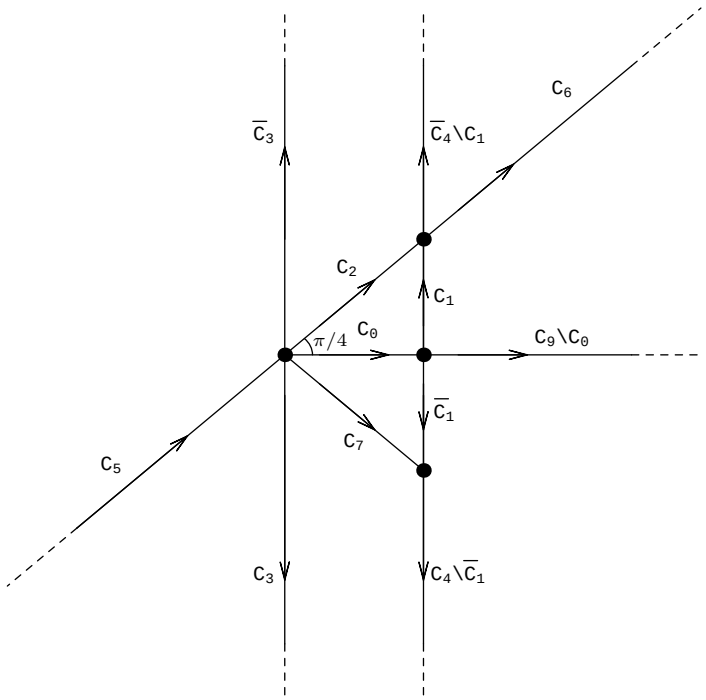

Figure 2: An illustration of the contours $C_0, C_1, \ldots, C_9$

Define the functions

$$\begin{aligned} I_C(K,j;a,b) &:= \frac{1}{K^j}\int_C t^j \exp(2\pi i a t + 2\pi i b t^2)\,dt \\ J(K,j;m,a,b) &:= \frac{1}{K^j}\int_0^K t^j \exp\left(-2\pi a t - 2\pi i b t^2\right)\frac{1-\exp\left(-2\pi m t\right)}{\exp(2\pi t)-1}\,dt \end{aligned}$$

It will be convenient to let $\tilde{I}_C(K,j;a,b) := I_C(K,j;ia,-b)$. If $j=0$, then j is omitted from the list of arguments; for example, we let $I_C(K;a,b) := I_C(K,0;a,b)$, $J(K;p_1,a,b) = J(K,0;p_1,a,b)$, and so on.

Let $\lfloor x \rfloor$ denote the largest integer less than or equal to x , $\lceil x \rceil$ denote smallest integer greater than or equal to x, $\{x\}$ denote $x - \lfloor x \rfloor$, and $\log x$ denote $\log_e x$. With this notation, define

$$\begin{aligned} p &:= \lceil a \rceil, & q &:= \lfloor a + 2bK \rfloor \\ p_1 &:= q - p, & \omega &:= \{a + 2bK\} \\ \omega_1 &:= \lceil a \rceil - a, & \nu(K,l,\epsilon) &:= (l+1)\log K/\epsilon \end{aligned}$$

We let $\exp(x)$ and e^x stand for the usual exponential function (and are used interchangeably). Finally, we define $0^0 := 1$ whenever it occurs.

3 The Typical iteration for $F(K,j;a,b)$

An *arithmetic operation* means an addition, a multiplication, an evaluation of the logarithm of a positive number, or an evaluation of complex exponential. We measure computational complexity (or time) by the number of arithmetic operations required. This in turn can be routinely bounded in terms of bit operations because all the numbers that occur in our algorithm have $O(\nu(K,j,\epsilon)^2)$ bits. Frequently, we abbreviate "arithmetic operations" to simply "operations."

For any $j \geq 0$ and $\epsilon \in (0, e^{-1})$, we say $K > 0$ is *large enough* if it is satisfies the lower bound $K > \Lambda$ where $\Lambda(K,j,\epsilon) := 1000\nu(K,j,\epsilon)^6$. In particular, if K is large enough, then $e^{-K} = o((\epsilon/K)^{1000(j+1)})$. We also let $\nu(K,\epsilon) := \nu(K,0,\epsilon)$ and $\Lambda(K,\epsilon) := \Lambda(K,0,\epsilon)$.

We call a real pair (a,b) *normalized* if $(a,b) \in [0,1) \times [0,1/4]$. The normalization is important because sums are converted to integrals via Poisson summation. Therefore, different choices of a or b produce different integrals. We remark it is mainly the normalization of quadratic argument b that truly matters. Normalizing a so that it is in the interval $[0,1)$ is not critical to what follows; for example, it suffices to take $a \in [-m,m]$ for some $0 < m = O(1)$.

Let j be a non-negative integer, ϵ any number in the interval $(0,e^{-1})$, K any large enough integer, and (a,b) any normalized pair. Then, with p and q defined as in Section 2, either $q > p$ or $q \leq p$. The first possibility is the main case, and it is where the method typically spends most of its time. The second possibility is a boundary point. We show how to handle each case separately. Arithmetic is performed using $O(\nu(K,j,\epsilon)^2)$-bits.

3.1 Case: $q > p$

By Poisson summation we have

$$F(K,j;a,b) = \frac{1}{2}\left(\delta_j + \exp(2\pi iaK + 2\pi ibK^2)\right) + PV \sum_{m=-\infty}^{\infty} I_{C_0}(K,j;a-m,b)$$

where δ_j is Kronecker's delta, and PV stands for Principal Value. Define

$$S_1(K,j;a,b) := \sum_{m=p}^{q} I_{C_0}(K,j;a-m,b)$$
$$S_2(K,j;a,b) := PV \sum_{m\notin[p,q]} I_{C_0}(K,j;a-m,b)$$

By Lemma 6.1, the condition $q > p$ implies $2bK \geq 1$. This simple observation is used repeatedly throughout Section 3.1, often without any comment.

3.1.1 The sum $S_1(K,j;a,b)$

By Cauchy's Theorem

$$I_{C_0}(K,j;a-m,b) = I_{C_2}(K,j;a-m,b) - I_{C_1}(K,j;a-m,b) \qquad (10)$$

Consider the integral over C_1 first. Ignoring the term $m = q$ for now and summing over $m \in \{p,\ldots,q-1\}$ we obtain

$$\sum_{m=p}^{q-1} I_{C_1}(K,j;a-m,b) = c_1 \sum_{l=0}^{j} i^l \binom{j}{l} J(K,l;p_1,\omega,b)$$

where $c_1 = i\exp(2\pi iaK + 2\pi ibK^2)$. When $m = q$, Cauchy's Theorem yields

$$\begin{aligned} I_{C_1}(K,j;a-q,b) &= c_1 \sum_{l=0}^{j} i^l \binom{j}{l} \tilde{I}_{C_0}(K,l;\omega,b) \\ &= c_1 \sum_{l=0}^{j} i^l \binom{j}{l} \tilde{I}_{C_7}(K,l;\omega,b) - c_1 \sum_{l=0}^{j} i^l \binom{j}{l} \tilde{I}_{\overline{C_1}}(K,l;\omega,b) \end{aligned}$$

According to lemmas 6.2 and 6.3, there exist absolute constants $\tilde{\kappa}_3$and $\tilde{\kappa}_4$ such that for any $\epsilon \in (0,e^{-1})$, any integer $l \geq 0$, any integer $M = O(e^{2K})$, and any integer $K > \Lambda(K,l,\epsilon)$, the functions $J(K,l;M,\omega,b)$ and $\tilde{I}_\gamma(K,l;\omega,b)$, where $\gamma \in \{C_7, \overline{C_1}\}$, can be computed in $O(\nu(K,j,\epsilon)^{\tilde{\kappa}_3})$ operations with error bounded by $O(\nu(K,j,\epsilon)^{\tilde{\kappa}_4} 8^{-j} K^{-2}\epsilon)$.

Having disposed of the integral over C_1 in (10), we now consider the one over C_2. Write

$$I_{C_2}(K,j;a-m,b) = I_{C_8}(K,j;a-m,b) - I_{C_5}(K,j;a-m,b) - I_{C_6}(K,j;a-m,b)$$

Consider C_5 first. Cauchy's theorem and a straightforward estimate yield

$$\sum_{m=p+1}^{q} I_{C_5}(K,j;a-m,b) = c_2 J\,(K,j;p_1,\omega_1,b) + O(e^{-K})$$

where $c_2 = (-1)^j e^{(j+1)\pi i/2}$. When $m = p$ we have

$$I_{C_5}(K,j;a-p,b) = c_2 \tilde{I}_{C_7}(K,j;\omega_1,b) + O(e^{-K})$$

As before, the integrals $J(.)$ and $\tilde{I}(.)$ in the last two equalities are handled by lemmas 6.2 and 6.3. As for C_6, it is not hard to see that $I_{C_6}(K,j;a-m,b)$ is $O(e^{-K}/K)$ for each $m = p,\ldots,q-1$; hence, these terms are negligible. When $m = q$, we obtain

$$I_{C_6}(K,j;a-q,b) = c_3 e^{-2\pi\omega K} \sum_{l=0}^{j} \binom{j}{l} 2^{\frac{j+1}{2}} \tilde{I}_{C_9}(K,l;\omega - i\omega + 2bK + i2bK, -2ib)$$

where $c_3 = e^{(j+1)\pi i/4 + 2\pi i a K}$. The integrals $\tilde{I}_{C_9}(K,l;\omega - i\omega + 2bK + i2bK, -2ib)$ are handled by Lemma 6.3. Finally, by Lemma 6.4

$$\sum_{m=p}^{q} I_{C_8}(K,j;a-m,b) = \sum_{s=0}^{j} w_{s,j} F(q,s;a_1,b_1) - \delta_{1-p} \sum_{s=0}^{j} w_{s,j} \qquad (11)$$

where $a_1 \equiv a/(2b) (\mathrm{mod}\ 1)$, $b_1 \equiv -1/(4b) (\mathrm{mod}\ 1)$, and all $w_{s,j}$ can be computed using $O(j^4)$ operations.

For purposes of our error analysis, it is sufficient to find the maximum modulus of $w_{s,j}$ during a full run of the algorithm. Suppose $2bK > \Lambda(K,j,\epsilon)$, then by Lemma 6.5

$$\sum_{m=s}^{j} |w_{s,m}| \le \frac{e^{1+j/\Lambda}}{\sqrt{2b}}(1 + 2j/\Lambda) \le \frac{e^{1+1/\log_2 K}}{\sqrt{2b}}(1 + 1/\log_2 K)$$

Now, suppose $0 < 2bK \le \Lambda(K,j,\epsilon)$. This can happen only in the last iteration. There are two possibilities: either $p < q \le \Lambda(K,j,\epsilon) + 1$ or $q \le p$. In the latter case we do not reach (11) at all. In the former case, Lemma 6.5 gives $\sum_{m=s}^{j} |w_{s,m}| \le (j+1)4^{j+2}(2b)^{-1/2}$. Recall (9). Then, put together, the maximum modulus of the coefficients $w_{s,j}$ that can occur over a full run of the algorithm satisfies the bound $O((j+1)4^j\sqrt{K}e^{\log_2 K}) = O((j+1)4^j K^2)$. This concludes our computation of $S_1(K,j;a,b)$.

3.1.2 The sum $S_2(K,j;a,b)$

Let us first handle the subsum $\sum_{m=q+1}^{M} I_{C_0}(K,j;a-m,b)$, $M > q$. For $m \ge q+1$

$$|I_{C_0 - iT}(K,j;a-m,b)| \le (2T)^j e^{-2\pi(1-\omega)T} \int_0^K e^{-4\pi bT(K-t)}\,dt \to_{T\to\infty} 0 \quad (12)$$

So, Cauchy's Theorem gives

$$I_{C_0}(K,j;a-m,b) = I_{C_3}(K,j;a-m,b) - I_{C_4}(K,j;a-m,b) \qquad (13)$$

We remark if $j=0$, then (12) holds uniformly in $a \in [0,2]$ and $m > q$. Therefore, (13) holds for all $a \in [0,2]$ and $m > q$. This observation is used the proof of Lemma 6.7. By a simple calculation

$$\sum_{m=q+1}^{M} I_{C_3}(K,j;a-m,b) = c_4 J(K,j;M-q,2bK-\omega,b) + O(e^{-K})$$

where $c_4 = (-i)^{j+1}$. A similar calculation gives

$$\sum_{m=q+2}^{M} I_{C_4}(K,j;a-m,b) = c_5 \sum_{l=0}^{j} (-i)^l \binom{j}{l} J(K,l;M-q-1,1-\omega,b) + O(e^{-K})$$

with $c_5 = -ie^{2\pi iaK+2\pi ibK^2}$. Furthermore

$$I_{C_4}(K,j;a-q-1,b) = c_5 \sum_{l=0}^{j} (-i)^l \binom{j}{l} \tilde{I}_{C_9}(K,l;1-\omega,b)$$

The integrals $\tilde{I}_{C_9}(K,l;1-\omega,b)$ are handled by Lemma 6.3.

As for $\sum_{m=-M}^{p-1} I_{C_0}(K,j;a-m,b)$, the situation is analogous. Simply use the conjugates of the contours C_3 and C_4, then repeat the same calculations with the appropriate modifications. The resulting integrals are

$$\sum_{m=-M}^{p-2} I_{\overline{C_3}}(K,j;a-m,b) = c_6 J(K,j;M+p-1,1-\omega_1,b)$$

$$\sum_{m=-M}^{p-1} I_{\overline{C_4}}(K,j;a-m,b) = c_7 \sum_{l=0}^{j} \binom{j}{l} i^l J(K,l;M+p,2bK-\omega_1,b)$$

$$I_{\overline{C_3}}(K,j;a-p+1,b) = c_6 \tilde{I}_{C_9}(K,j;1-\omega_1,b)$$

where $c_6 := i^{j+1}$ and $c_7 := ie^{2\pi iaK+2\pi ibK^2}$. Finally, the terms corresponding to $|m| > M$ can be bounded as follows. Write

$$PV \sum_{|m|>M} I_{C_0}(K,j;a-m,b) = \sum_{m>M} \frac{2}{K^j} \int_0^K x^j \exp(2\pi iax + 2\pi ibx^2) \cos(2\pi mx)\, dx$$

Integrating by parts this is equal to

$$-\sum_{m>M} \left(\frac{j}{\pi m K^j} \int_0^K (1-\delta_j) x^{j-1} \exp(2\pi iax + 2\pi ibx^2) \sin(2\pi mx)\, dx + \right.$$

$$\left. \frac{2i}{mK^j} \int_0^K x^j\, (a+2bx) \exp(2\pi iax + 2\pi ibx^2) \sin(2\pi mx)\, dx \right)$$

By the Second Mean Value Theorem, we deduce for $M > 2K$

$$PV \sum_{|m|>M} I_{C_0}(K, j; a-m, b) = O\left(\sum_{m>M} \frac{K}{m(m-K)}\right) = O\left(\frac{K}{M}\right)$$

Finally, take $M = e^{2K}$ to obtain the bound $O(8^{-j}K^{-2}e^{-K})$.

3.2 Case: $q \leq p$

The case $q \leq p$ is a terminal point of the method. Observe if $q \leq p$, then $0 \leq a + 2bK < 2$. Thus, we may assume $1/K^2 < b < 1/K$ (if $b \leq 1/K^2$, then the problem is trivial). We may also assume K is a multiple of 8. Write

$$F(K, j; a, b) = \frac{e^{2\pi i a K + 2\pi i b K^2}}{K^j} + \frac{1}{K^j} \sum_{m=0}^{7} \sum_{k=mK/8}^{(m+1)K/8-1} k^j \exp(2\pi i a k + 2\pi i b k^2)$$

But

$$\frac{1}{K^j} \sum_{k=mK/8}^{(m+1)K/8} k^j \exp(2\pi i a k + 2\pi i b k^2) = c_{13}\, 8^{-j} \sum_{l=0}^{j} m^{j-l} \binom{j}{l} F(K_1, l; \tilde{a}, b)$$

where $|c_{13}| = 1$, $0 \leq m < 8$, $K_1 := K/8$, and $\tilde{a} := a + mbK/4$. We can normalize so that $-1/2 \leq \tilde{a} \leq 1/2$. Since $0 \leq 2bK < 2$, then $0 \leq 2bK_1 < 1/4$. So, $0 \leq |\tilde{a}| + 2bK_1 < 3/4$. Put together, we may now assume that $|a| + |2bK| < 3/4$, where $a, b \in \mathbb{R}$. Define the function

$$f_\alpha(x) := \frac{x^\alpha}{K^\alpha} \exp(2\pi i a x + 2\pi i b x^2)$$

where α is a non-negative integer. By Lemma 6.6

$$\max_{0 \leq x \leq K} |f_\alpha^{(N)}(x)| \leq \left(\frac{\alpha + N}{K} + 2\pi(|a| + |2bK|)\right)^N$$

So, given a sum $F(K, j; a, b)$ and $\epsilon \in (0, e^{-1})$ such that $K > \Lambda(K, j, \epsilon)$, $a, b \in \mathbb{R}$ and $|a| + |2bK| < 3/4$, we can apply Euler-Maclaurin Summation to $F(K, j; a, b)$ with N correction terms. Take $N = \log(8^j K^3/\epsilon)/(2\log(8/7))$. Then, resulting absolute error is bounded by

$$\frac{2}{(2\pi)^{2N}} \int_0^K |f_j^{(2N+1)}(x)|\, dx \leq 2K(7/8)^{-2N} = O(8^{-j}K^{-2}\epsilon)$$

It remains to evaluate the integral $I_{C_0}(K, j; a, b)$. But this is handled by Lemma 6.3.

4 The Algorithm for $F(K, j; a, b)$

Lemma 4.1. *For any integer $K \geq 0$, any integer $j \geq 0$, and any $a, b \in \mathbb{C}$, the function $F(K, j; a, b)$ satisfies the identities*

$$F(K, j; a, b) = F(K, j; a + 1, b) = F(K, j; a, b + 1) = F(K, j; a \pm 1/2, b \pm 1/2)$$

Proof. This follows from $e^{2\pi i(z+1)} = e^{2\pi iz}$ and $(k^2 \pm k)/2 \in \mathbb{Z}$ for $k \in \mathbb{Z}$. □

Lemma 4.1 implies that for any integers $K \geq 0$ and $j \geq 0$, there exists a pair $(a_0, b_0) \in [0, 1) \times [0, 1/4]$ such that $F(K, j; a, b)$ is equal to $F(K, j; a_0, b_0)$ or its conjugate (this is independent of K and j). We are now ready to present pseudo-code to compute $\sum_{l=0}^{j} z_{l,j} F(K, l; a, b)$; $z_{l,j} = O(1)$. It suffices to do the arithmetic using $O(\nu(K, j, \epsilon)^2)$ bits.

INPUT: $a, b \in \mathbb{R}/\mathbb{Z}$, an integer $K > 0$, a positive number $\epsilon < e^{-1}$, an integer $j \geq 0$, and an array of complex numbers $z_{l,j}$, $l = 0, \ldots, j$ such that $z_{l,j} = O(1)$.

OUTPUT: a complex number *sum* that equals $\sum_{l=0}^{j} z_{l,j} F(K, l; a, b)$ modulo error $O(\nu(K, j, \epsilon)^{\tilde{\kappa}_1 + 1} \epsilon)$, where $\tilde{\kappa}_1$ is an absolute constant.

INITIALIZE: $sum = 0$, $flag = 0$, $counter = 0$

1. Normalize $(a, b) \to (a_0, b_0)$. Set $flag = 1$, $z_{l,j} \to \overline{z_{l,j}}$ if conjugation is needed to normalize (a, b) (see Lemma 4.1).

2. Set $p = \lceil a_0 \rceil$, $q = \lfloor a_0 + 2b_0 K \rfloor$.

3. If K is large enough, then go to 4. Otherwise, compute the sum directly, let $R[counter]$ be the result. If $flag = 1$, set $R[counter] \to \overline{R[counter]}$. Go to 9.

4. If $q \leq p$, then apply the typical iteration in this case, let $R[counter]$ denote the result. If $flag = 1$, set $R[counter] \to \overline{R[counter]}$. Go to 9.

5. Apply the typical iteration in the case $q > p$

$$F(K, j; a_0, b_0) \to \sum_{l=0}^{j} w_{l,j} F(K_1, l; a_1, b_1) + R_j + O(\nu(K, j, \epsilon)^{\tilde{\kappa}_1} \epsilon)$$

$$F(K, j-1; a_0, b_0) \to \sum_{l=0}^{j-1} w_{l,j-1} F(K_1, l; a_1, b_1) + R_{j-1} + O(\nu(K, j, \epsilon)^{\tilde{\kappa}_1} \epsilon)$$

$$\vdots$$

$$F(K, 0; a_0, b_0) \to w_{0,0} F(K_1, l; a_1, b_1) + R_0 + O(\nu(K, j, \epsilon)^{\tilde{\kappa}_1} \epsilon)$$

6. Set $R[counter] = \sum_{l=0}^{j} z_{l,j} R_l$, $z_{l,j} \to \sum_{s=l}^{j} z_{s,j} w_{l,s}$, $K \to K_1 = \lfloor a_0 + 2b_0 K \rfloor$, $counter \to counter + 1$, $a \to a_1$, $b \to b_1$.

7. If $flag = 1$, then set $z_{l,j} \to \overline{z_{l,j}}$, $R[counter] \to \overline{R[counter]}$, $a \to -a$, $b \to -b$, $flag \to 0$.

8. Goto 1.

9. $sum \to \sum_{l=0}^{counter} R[l]$.

5 The Sums $G(K, j; a, b)$

We show how evaluate the sums $G(K, j; a, b)$. Let N be a positive multiple of 16. Define

$$V(N, j; a, b) := \sum_{k=N}^{2N-1} \frac{1}{k^j} \exp(2\pi i a k + 2\pi i b k^2)$$

Since $G(K, j; a, b)$ consists of $O(\log K)$ sums of the form $V(N, j; a, b)$ plus a remainder sum of length $O(\log K)$, it is enough to compute V. We can write

$$V(N, j; a, b) = \sum_{m=0}^{15} \sum_{k=N+mN/16}^{N+(m+1)N/16-1} \frac{1}{k^j} \exp(2\pi i a k + 2\pi i b k^2)$$

Define $N_m := N + mN/16$. Then

$$\sum_{k=N_m}^{N_{m+1}-1} \frac{1}{k^j} \exp(2\pi i a k + 2\pi i b k^2) \tag{14}$$

$$= \frac{c_{1,m}}{N_m^j} \sum_{l=0}^{\infty} (-1)^l \binom{j+l-1}{j-1} \sum_{k=0}^{N/16-1} \frac{k^l}{N_m^l} \exp(2\pi i \tilde{a} k + 2\pi i b k^2)$$

where $|c_{1,m}| = 1$. Since $\binom{j+l-1}{j-1} k^l / N_m^{l+j} \le 8^{-l}$, we can truncate the sum over l after $\nu(K, j, \epsilon)$. Finally, the inner sums in (14) are handled by Theorem 1.1.

6 Auxiliary Results

Lemma 6.1. *Let $a, b \in \mathbb{R}$ and $K \ge 0$. If $\lfloor a + 2bK \rfloor > \lceil a \rceil$ then $2bK \ge 1$.*

Proof. $\lfloor a + 2bK \rfloor > \lceil a \rceil \Rightarrow 2bK + a \ge \lceil a \rceil + 1$. So, $2bK \ge 1$. □

Lemma 6.2. *Let $J(.)$, $\nu(.)$, and $\Lambda(.)$ be defined as in Section 2. There exist absolute constants κ_3 and κ_4 such for any $\epsilon \in (0, e^{-1})$, any integer $l \ge 0$, any positive integer K satisfying $K > \Lambda(K, l, \epsilon)$, any $b \in [0, 1]$ satisfying $2bK \ge 1$, any $0 \le \omega \le 1$, any $w \ge 0$, and any positive integer M satisfying $M = O(e^{2K})$, the integral $J(K, l; M, w + \omega, b)$ can be evaluated with error bounded by $O(\nu(K, l, \epsilon)^{\kappa_3} 8^{-l} K^{-2} \epsilon)$ using $O(\nu(K, l, \epsilon)^{\kappa_4})$ operations. Big-O constants are absolute and arithmetic is done using $O(\nu(K, l, \epsilon)^2)$ bits.*

Proof. Let us assume $w = 0$, the case $w > 0$ is similar. Let $L = \lceil 3\nu(K, l, \epsilon) \rceil$. Then

$$\begin{aligned} J(K, l; M, \omega, b) &= \frac{1}{K^l} \int_0^L t^l \exp\left(-2\pi\omega t - 2\pi i b t^2\right) \frac{1 - \exp\left(-2\pi M t\right)}{\exp(2\pi t) - 1}\, dt \\ &\quad + O(8^{-l} K^{-3} \epsilon) \end{aligned}$$

So, it suffices to evaluate the integrals

$$\frac{1}{K^l} \int_n^{n+1} t^l \exp\left(-2\pi\omega t - 2\pi i b t^2\right) \frac{1 - \exp\left(-2\pi M t\right)}{\exp(2\pi t) - 1}\, dt$$

where $n = 0, \ldots, L-1$. Taylor expansions yield

$$\begin{aligned} &\frac{\exp(-2\pi\omega n - 2\pi i b n^2)}{K^l} \sum_{s=0}^{l} \binom{l}{s} n^{l-s} \sum_{r=0}^{L} \frac{(-2\pi i b)^r}{r!} \times \\ &\int_0^1 t^{s+2r} \exp\left(-2\pi\omega t - 4\pi i b n t\right) \frac{1 - \exp\left(-2\pi M (t+n)\right)}{\exp(2\pi(t+n)) - 1}\, dt + O((\log M) 8^{-l} K^{-3} \epsilon) \end{aligned}$$

In order to evaluate the last expression, it suffices to deal with the integrals

$$\int_0^1 t^\alpha \exp\left(-2\pi\omega t - 4\pi i b n t\right) \frac{1 - \exp\left(-2\pi M (t+n)\right)}{\exp(2\pi(t+n)) - 1}\, dt \tag{15}$$

for integers $\alpha \in [0, 3L]$. These can be evaluated by unfolding the geometric series; that is, write (15) as

$$\sum_{m=1}^{M} \exp(-2\pi m n) \int_0^1 t^\alpha \exp\left(-2\pi(m + \omega + 2ibn)t\right)\, dt \tag{16}$$

Apply Taylor expansions to the factor $\exp(-2\pi\omega t)$ in (16) to reduce its evaluation to that of sums of the form

$$\sum_{m=1}^{M} \exp(-2\pi m n) \int_0^1 t^{\alpha+\beta} \exp\left(-2\pi(m + 2ibn)t\right)\, dt \tag{17}$$

for integers $0 \le \beta \le L$. Let $\alpha_1 := \alpha + \beta$. For $m \ge \alpha_1 + 1$, we explicitly evaluate (17) to obtain

$$\begin{aligned} &\sum_{v=1}^{\alpha_1+1} (-1)^{v+1} \frac{\alpha_1!}{(\alpha_1 + 1 - v)!} \sum_{m=\alpha_1+1}^{M} \exp(-2\pi m n) \frac{\exp(-2\pi m - 4\pi b i n))}{(2\pi m + 4\pi i b n)^v} \\ &+ (-1)^{\alpha_1} \alpha_1! \sum_{m=\alpha_1+1}^{M} \frac{\exp(-2\pi m n)}{(2\pi m + 4\pi i b n)^{\alpha_1+1}} \end{aligned} \tag{18}$$

The sums in (18) can be evaluated by truncation or Euler-Maclaurin summation (in the case $n = 0$ in the second sum). Now, consider the terms

$1 \le m \le \alpha_1$. For each integer $0 \le n < L$ and each integer $1 \le m \le \alpha_1$, there are two possibilities: either $n+1 \le m \le \alpha_1$ or $1 \le m < n+1$. In the first case, apply the change of variable $t \to mt$ to the integral in (17) to reduce it to integrals of the form

$$\frac{1}{m^{\alpha_1}} \int_l^{l+1} t^{\alpha_1} \exp\left(-2\pi(1+2ibn/m)t\right)\, dt$$

where l is an integer in $[0, m]$. These integrals are straightforward to evaluate; first, make a change of variable $t \to t - l$, then use Taylor expansions to reduce the integrand to a polynomial in t of degree $O(L)$. Finally, the case $1 \le m < n+1$ is handled analogously. □

Lemma 6.3. *Let $I_C(.)$, $\tilde{I}_C(.)$, $\nu(.)$, $\Lambda(.)$, C_1, C_7, and C_9 be defined as in Section 2. There exist absolute constants κ_5 and κ_6 such for any $\epsilon \in (0, e^{-1})$, any integer $j \ge 0$, any positive integer K satisfying $K > \Lambda(K, j, \epsilon)$, any $b \in [0, 1]$ satisfying $2bK \ge 1$, any real a satisfying $a = O(1)$, any $0 \le \omega \le 1$, and any $0 < \delta < 1$, each of the integrals*

$$\tilde{I}_{\overline{C_1}}(K, j; \omega, b), \qquad \tilde{I}_{C_7}(K, j; \omega, b)$$
$$\tilde{I}_{C_9}(K, j; \delta + \omega, b), \qquad \tilde{I}_{C_9}(K, j; \omega - i\omega + 2bK + i2bK, -2ib)$$

can be evaluated with error bounded by $O(\nu(K, j, \epsilon)^{\kappa_5} 8^{-l} K^{-2} \epsilon)$ using $O(\nu(K, j, \epsilon)^{\kappa_6})$ operations. Under the same assumptions, except now b need not satisfy $2bK \ge 1$, the integral $I_{C_0}(K, j; a, b)$ can be evaluated with the same error and complexity. Big-O constants are absolute and arithmetic is done using $O(\nu(K, l, \epsilon)^2)$ bits.

Proof. We evaluate $\tilde{I}_{\overline{C_1}}(K, j; \omega, b)$ first. We have

$$\tilde{I}_{\overline{C_1}}(K, j; \omega, b) = c_8 e^{-2\pi\omega K} \sum_{l=0}^{j} \binom{j}{l} \frac{(-i)^l}{K^l} \int_0^K t^l \exp\left(2\pi i\omega t - 4\pi bKt + 2\pi ibt^2\right)\, dt$$

where $c_8 := -i \exp(-2\pi ibK^2)$. By truncating at $L = \lceil 3\nu(K, j, \epsilon) \rceil$, the evaluation of $\tilde{I}_{\overline{C_1}}(K, j; \omega, b)$ is reduced to that of integrals of the form

$$\frac{1}{L^l} \int_n^{n+1} t^l \exp\left(2\pi i\omega t - 4\pi bKt + 2\pi ibt^2\right)\, dt \tag{19}$$

for integers $0 \le l \le j$ and $0 \le n \le L - 1$. In order to compute (19), substitute $t \to t - n$, then eliminate the quadratic term $2\pi ibt^2$ using Taylor expansion. This results in easily-calculable integrals of the form

$$\int_0^1 t^\alpha \exp\left(2\pi i(\omega + 2bn)t - 4\pi bKt\right)\, dt$$

where α is an integer in $[0, 3L]$. The evaluation of

$$\tilde{I}_{C_9}(K, l; \omega - i\omega + 2bK + i2bK, -2ib)$$

is similar to $\tilde{I}_{\overline{C_1}}(K,j;\omega,b)$. So, we may move on to $\tilde{I}_{C_7}(K,j;\omega,b)$. By definition

$$\tilde{I}_{C_7}(K,j;\omega,b) \quad = \quad \frac{c_9}{K^j}\int_0^{\sqrt{2}K} t^j \exp\left(-\sqrt{2}\pi\omega t + \sqrt{2}\pi i\omega t - 2\pi b t^2\right)\,dt$$

where $c_9 = \exp\left(-(j+1)\pi i/4\right)$. The change of variable $t \to \sqrt{b}t$ yields

$$\tilde{I}_{C_7}(K,j;\omega,b) \quad = \quad \frac{c_9}{b^{(j+1)/2}K^j}\int_0^{\sqrt{2b}K} t^j \exp\left(-2\pi\frac{\omega}{\sqrt{2b}}t + 2\pi i\frac{\omega}{\sqrt{2b}}t - 2\pi t^2\right)\,dt$$

So, truncating the integral at $\lceil\sqrt{L}\rceil$ reduces the problem to evaluating

$$\frac{1}{b^{(j+1)/2}K^j}\int_n^{n+1} t^j \exp\left(-2\pi\frac{\omega}{\sqrt{2b}}t + 2\pi i\frac{\omega}{\sqrt{2b}}t - 2\pi t^2\right)\,dt$$

for integers $0 \le n < \lceil\sqrt{L}\rceil$. Finally, these integrals are handled as before; substitute $t \to t-n$, then eliminate the quadratic term using Taylor expansions. This procedure results in readily-calculable integrals. Next, we consider $\tilde{I}_{C_9}(K,j;\delta+\omega,b)$. Let $\omega' := \omega + \delta$. Then $0 < \omega' < 2$ and

$$\left|\frac{1}{K^j}\int_0^T (T-it)^j \exp\left(-2\pi\omega'(T-it) - 2\pi i b(T-it)^2\right)\,dt\right| \to_{T\to\infty} 0 \qquad (20)$$

So, we can replace C_9 by $e^{-\pi i/4}C_9$ in $\tilde{I}_{C_9}(K,j;\omega',b)$. A simple estimate then yields

$$\tilde{I}_{e^{-\pi i/4}C_9}(K,j;\omega',b) = \tilde{I}_{C_7}(K,j;\omega',b) + O(e^{-K}) \qquad (21)$$

We have already shown how to compute the right side of (21). We remark if $j=0$, then (20), hence (21), holds uniformly for $\omega' \in [0,1]$. This observation is used in the proof of Lemma 6.7.

Finally, we consider the integral $I_{C_0}(K,j;a,b)$. This may contain a critical point or it may not according to whether $-a/(2b) \in [0,K]$ or not. We showed how to deal with these possibilities in Sections 3.1.1 and 3.1.2 respectively. □

Lemma 6.4. *Let $I_C(.)$ and C_8 be defined as in Section 2. For any integer $K > 0$, any integer $j \ge 0$, any integer m, any $a \in \mathbb{R}$, and any $b > 0$ such that $q := \lfloor a + 2bK \rfloor$ is not zero, we have*

$$I_{C_8}(K,j;a-m,b) = \exp\left(\frac{2\pi i a}{2b}m - \frac{2\pi i}{4b}m^2\right)\sum_{s=0}^{j}\frac{w_{s,j}m^s}{q^s} \qquad (22)$$

where

$$w_{s,j} = q^s\frac{j!\,\sqrt{2\pi}e^{\pi i/4}e^{(j-s)3\pi i/4}e^{-i\pi a^2/(2b)}}{2^{j/2}s!(2\sqrt{b\pi})^{j+1}K^j}\left(\sqrt{\frac{2\pi}{b}}\right)^s \times$$
$$\sum_{l=0}^{j-s}\frac{\delta_{(j-s-l)(mod\,2)}(-1)^{(j+l-s)/2}}{l!\frac{j-s-l}{2}!}\left(ae^{-3\pi i/4}\sqrt{\frac{2\pi}{b}}\right)^l \qquad (23)$$

We remark that (22) is what one would expect; it is also essentially independent of K.

Proof. By a change of variable

$$\begin{aligned} I_{C_8}(K,j;a-m,b) &= \frac{e^{(j+1)\pi i/4}}{(\sqrt{2}\pi)^{j+1}K^j}\int_{-\infty}^{\infty} t^j \exp(iat - at - bt^2/\pi - imt + mt)\,dt \\ &= \frac{e^{(j+1)\pi i/4}}{(\sqrt{2}\pi)^{j+1}K^j}\int_{-\infty}^{\infty} t^j \exp\left(\frac{i\sqrt{2b}\,\tilde{\alpha}}{\sqrt{\pi}}t - \frac{b}{\pi}t^2\right)dt \qquad (24) \end{aligned}$$

where

$$\tilde{\alpha} := \frac{\sqrt{\pi}(m-a)(1-i)}{i\sqrt{2b}} = \frac{\sqrt{\pi}e^{-3\pi i/4}(m-a)}{\sqrt{b}}$$

Let $H_j(x)$ denote the Hermite polynomials, which are defined by generating function

$$H_j(\alpha) := (-1)^j \exp\left(\alpha^2/2\right)\frac{d^j}{d\alpha^j}\exp\left(-\alpha^2/2\right)$$

By the absolute convergence of (24), it is possible to differentiate under the integral sign. Thus, we obtain

$$\begin{aligned} I_{C_8}(K,j;a-m,b) &= \frac{e^{(j+1)\pi i/4}}{(\sqrt{2}\pi)^{j+1}K^j}\left(\frac{\sqrt{\pi}}{i\sqrt{2b}}\right)^j \frac{d^j}{d\alpha^j}\int_{-\infty}^{\infty}\exp\left(\frac{i\sqrt{2b}}{\sqrt{\pi}}\alpha t - bt^2/\pi\right)dt\Bigg|_{\alpha=\tilde{\alpha}} \\ &= \frac{\sqrt{2\pi}e^{\pi i/4+3\pi ij/4}}{(2\sqrt{b\pi})^{j+1}K^j}\exp\left(-\tilde{\alpha}^2/2\right)H_j\left(\tilde{\alpha}\right) \end{aligned}$$

The coefficient of x^l in $H_j(x)$ can be interpreted as $(-1)^{(j-l)/2}$ times the number of unordered partitions of a j-element set into l singletons and $(j-l)/2$ unordered pairs (see [I]). Therefore,

$$I_{C_8}(K,j;a-m,b) = \frac{\sqrt{2\pi}e^{\pi i/4+3\pi ij/4}e^{-i\pi a^2/(2b)}}{(2\sqrt{b\pi})^{j+1}K^j}\exp\left(\frac{2\pi ia}{2b}m - \frac{2\pi i}{4b}m^2\right)\times$$
$$\sum_{l=0}^{j}\delta_{(j-l)(\mathrm{mod}\,2)}(-2)^{(l-j)/2}\left(\frac{j-l}{2}\right)!\binom{j}{l}\binom{j-l}{\frac{j-l}{2}}\left(\sqrt{\frac{\pi}{b}}\right)^l e^{-3\pi il/4}(m-a)^l$$

The sum over l can be rewritten as

$$\sum_{l=0}^{j}\sum_{s=0}^{l}\delta_{(j-l)(\mathrm{mod}\,2)}\frac{(-1)^{s+(l+j)/2}2^{-j/2}\,j!}{s!\,(l-s)!\,\frac{j-l}{2}!}\left(\sqrt{\frac{2\pi}{b}}\right)^l e^{-3l\pi i/4}a^{l-s}m^s \qquad (25)$$

Finally, change the order of summation in (25) to obtain the result. □

Lemma 6.5. *Let $\Lambda(.)$ and $\nu(.)$ be defined as in Section 2. For any $\epsilon \in (0, e^{-1})$, any $a \in [0,1]$, any $b \in [0,1]$, any integer $j \geq 0$, any positive*

integer K satisfying $K > \Lambda(K, j, \epsilon)$, any integer $0 \le s \le j$, and with $w_{s,m}$ defined as in (23) and under the condition $\lceil a \rceil < \lfloor a + 2bK \rfloor$, we have the bound

$$\sum_{m=s}^{j} |w_{s,m}| \le \frac{e}{\sqrt{2b}} \left(1 + \frac{1}{2bK}\right)^{j} \sum_{g=0}^{j} \left(\frac{j}{2bK}\right)^{g}$$

Moreover, if $2bK \le \nu(K, j, \epsilon)$, then $\sum_{m=s}^{j} |w_{s,m}| \le (j+1)4^{j+1}(2b)^{-1/2}$.

Proof. To obtain the first bound note that

$$\begin{aligned} \sum_{m=s}^{j} |w_{s,m}| &\le \frac{q^s}{(2bK)^s \sqrt{2b}} \sum_{m=0}^{j-s} \frac{(m+s)^m}{\left(\sqrt{2\pi}\right)^m (2bK)^m} \sum_{l=0}^{m} \delta_{(m-l)(\mathrm{mod}\,2)} \frac{\left(\sqrt{2\pi}|a|\right)^l b^{(m-l)/2}}{l!\, \frac{m-l}{2}!} \\ &\le \frac{e^{|a|}}{\sqrt{2b}} \left(1 + \frac{|a|}{2bK}\right)^{j} \sum_{m=0}^{j} \left(\frac{j}{2bK}\right)^{m} \end{aligned}$$

If $2bK \le \nu(K, j, \epsilon)$, then since $K > \Lambda(K, j, \epsilon)$, it follows that $b < 1/(2j+2)^2$. Therefore

$$\sum_{m=s}^{j} |w_{s,m}| \le \frac{2q^s}{(2bK)^s \sqrt{2b}} \sum_{m=0}^{j-s} \frac{(m+s)!}{s!\, m!\, (2bK)^m} \le \frac{(j+1)4^{j+1}}{\sqrt{2b}}$$

□

Lemma 6.6. *For any integer $\alpha \ge 0$, any integer $m \ge 0$, any integer $K > 0$, and any real numbers a and b, the function*

$$f_\alpha(x) := \frac{x^\alpha}{K^\alpha} \exp(2\pi i a x + 2\pi i b x^2)$$

satisfies

$$\max_{0 \le x \le K} |f_\alpha^{(m)}(x)| \le (2\pi(|a| + |2bK|) + (m+\alpha)/K)^m$$

Proof. Since a, b, and K are fixed, we suppress dependencies on them. We can write

$$f_\alpha^{(m)}(x) = P_m(x) \exp(2\pi i a x + 2\pi i b x^2)$$

where

$$P_m(x) := \sum_{l=0}^{m+\alpha} d_{l,m,\alpha} x^l$$

for some complex coefficients $d_{l,m,\alpha}$. Define $|P_m(x)|_1 := \sum_{l=0}^{m+\alpha} |d_{l,m,\alpha} x^l|$. Suppose

$$\max_{0 \le x \le K} |P_m(x)|_1 \le (2\pi(|a| + |2bK|) + (m+\alpha)/K)^m$$

Then

$$\begin{aligned}\max_{0\le x\le K}|P_{m+1}(x)|_1 &\le \max_{0\le x\le K}|2\pi i a P_m(x)|_1+\max_{0\le x\le K}|4\pi i b x P_m(x)|_1\\ &\quad+\max_{0\le x\le K}|P_m'(x)|_1\\ &\le (2\pi(|a|+|2bK|)+(m+1+\alpha)/K)^{m+1}\end{aligned}$$

□

Lemma 6.7. *For any $a\in\mathbb{R}$, any $b>0$, and any $K>0$ let*

$$\begin{aligned}q_a=\lfloor a+2bK\rfloor,\qquad p_a=\lceil a\rceil,\qquad p_{1,a}=q_a-p_a\\ \omega_a=\{a+2bK\},\qquad \omega_{1,a}=p_a-a,\qquad M=e^{2K}\end{aligned}\tag{26}$$

and

$$\begin{aligned}c_{1,a}=\delta_{2-p_a},\qquad c_{2,a}=\delta_{q_a-K_1-1},\qquad c_{3,a}=\delta_{q_a-K_1-2}\\ f(x)=e^{2\pi i x K},\qquad g(x)=f(x)e^{-2\pi K\omega_x},\qquad h(x)=\frac{1}{\sqrt{2b}}e^{-i\pi x^2/(2b)}\end{aligned}\tag{27}$$

Also, let $F(K;a,b):=F(K,0;a,b)$, $K^*=\lfloor 2bK\rfloor$, $a^*=a/(2b)$, and $b^*=-1/(4b)$. Then for any $K>1000$ and any tuple (α,a,b) in $[0,1]\times[0,2]\times[0,1/4]$ satisfying

$$\begin{aligned}p_{a+\alpha x}<q_{a+\alpha x},\qquad &\text{for all } x\in[-1/4,1/4]\\ (a+\alpha x,b)\in(0,2)\times[0,1/4],\qquad &\text{for all } x\in[-1/4,1/4]\end{aligned}\tag{28}$$

we have

$$\begin{aligned}F(K;a+\alpha x,b) &= e^{i\pi/4}h(a+\alpha x)F\left(K^*;a^*+\frac{\alpha x}{2b},b^*\right)+R(K,a+\alpha x,b)\\ &\quad+O(e^{-K}),\qquad \text{for } x\in[-1/4,1/4],\end{aligned}$$

where $R(K,a+\alpha x,b)$ is a linear combination of the functions (defined in Section 2)

$$\begin{aligned}&J(K;p_{1,a+\alpha x},\omega_{1,a+\alpha x},b),\qquad f(\alpha x)J(K;p_{1,a+\alpha x},\omega_{a+\alpha x},b)\\ &J(K;M,2bK-\omega_{a+\alpha x},b),\qquad f(\alpha x)J(K;M,1-\omega_{a+\alpha x},b)\\ &J(K;M,1-\omega_{1,a+\alpha x},b),\qquad f(\alpha x)J(K;M,2bK-\omega_{1,a+\alpha x},b)\end{aligned}\tag{29}$$

as well as the functions (defined in Section 2)

$$\begin{aligned}&g(a+\alpha x)\tilde I_{C_0}(K;-i\omega_{a+\alpha x}+2bK,-b),\qquad g(a+\alpha x)\tilde I_{C_0}(K;e^{\pi i/4}(-i\omega_{a+\alpha x}+2bK),-ib)\\ &\tilde I_{C_7}(K;\omega_{1,a+\alpha x},b),\qquad f(\alpha x)\tilde I_{C_7}(K;\omega_{a+\alpha x},b)\\ &\tilde I_{C_7}(K;1-\omega_{1,a+\alpha x},b),\qquad f(\alpha x)\tilde I_{C_7}(K;1-\omega_{a+\alpha x},b)\end{aligned}\tag{30}$$

and the functions

$$\begin{aligned} &c_{2,a+\alpha x}h(a+\alpha x)e^{2\pi i\alpha x(K_1+1)/(2b)}, && c_{3,a+\alpha x}h(a+\alpha x)e^{2\pi i\alpha x(K_1+1)/(2b)} \\ &c_{3,a+\alpha x}h(a+\alpha x)e^{2\pi i\alpha x(K_1+2)/(2b)}, && c_{1,a+\alpha x}h(a+\alpha x)e^{2\pi i\alpha x/(2b)} \\ &f(\alpha x), && h(a+\alpha x) \end{aligned} \tag{31}$$

as well as the constant function. The coefficients in the linear combination can all be computed using $O(1)$ operations, are bounded by $O(1)$, and do not depend on x. Big-O constants are absolute.

Proof. This follows from our main result and the remarks following formulas (13) and (21). □

Lemma 6.8. *Let $\nu(K,\epsilon)$ and $\Lambda(K,\epsilon)$ be defined as at the beginning of Section 3. Recall the definitions (26), (27), and the definition $K^* := \lfloor 2bK \rfloor$ in Lemma 6.7. Then, for any $\epsilon \in (0, e^{-1})$, any K large enough, any interval $(w,z) \subset [-1/4, 1/4]$ and tuple*

$$(\alpha, a, b) \in [0, 1/\Lambda(K,\epsilon)] \times (0,2) \times [0, 1/4]$$

satisfying assumptions (28) as well as the assumption

$$p_{a+\alpha x} \text{ and } q_{a+\alpha x} \text{ are constant over } x \in (w,z),$$

and for $x \in (w,z)$, there exist constants $\lambda_\theta^+, \lambda_\theta^- \in [0,1]$, $\theta \in \{-1,1\}$, independent of x and satisfying $0 \le \lambda_\theta^\pm + \theta\alpha x \le 1$ for $x \in (w,z)$, such that each of the functions in (29), (30), and (31) is equal to a linear combination of functions of the form

$$\begin{aligned} &x^m, \qquad x^m e^{2\pi i\alpha K x}, \qquad e^{2\pi i\frac{\alpha M}{2b}x - 2\pi i\frac{\alpha^2}{4b}x^2}, \\ &(\lambda_\theta^\pm + \theta\alpha x)^m e^{2\pi i L(\lambda_\theta^\pm + \theta\alpha x) - 2\pi R(\lambda_\theta^\pm + \theta\alpha x)}, \\ &e^{2\pi i N(\lambda_\theta^\pm + \theta\alpha x) - 2\pi(1-i)m\frac{\lambda_\theta^\pm + \theta\alpha x}{\sqrt{2b}}} \int_0^1 t^l e^{-2\pi(1-i)\frac{\lambda_\theta^\pm + \theta\alpha x}{\sqrt{2b}}t - 2\pi mt}\, dt, \end{aligned}$$

plus an error bounded by $O(\Lambda(K,\epsilon)K^{-2}\epsilon)$. Here, $0 \le l, m = O(\nu(K,\epsilon))$ are integers and

$$\begin{aligned} N \in \{0, K\}, \qquad & M \in \{-1, 0, K^*, K^*+1\}, \\ K \le L \le K+u, \qquad & K \le R \le K+v \\ 0 \le u = O(\nu(K,\epsilon)) \qquad & 0 \le v = O(\nu(K,\epsilon)) \end{aligned}$$

The length of the linear combination is $O(\Lambda(K,\epsilon))$. The coefficients in the linear combinations can all be computed in $O(\Lambda(K,\epsilon))$ operations, are bounded by $O(K)$, and are independent of x. Big-O constants are absolute and arithmetic is done using $O(\nu(K,\epsilon)^2)$ bits.

Proof. This is readily deducible from the proofs of lemma 6.2, lemma 6.3, and the assumption that $p_{a+\alpha x}$ and $q_{a+\alpha x}$ are constant over (w,z). □

Acknowledgment

I am thankful to advisor Andrew Odlyzko. Without his extensive help and many comments this paper would not have been possible.

References

[D] Harold Davenport, *Multiplicative number theory*. Third edition. Revised and with a preface by Hugh L. Montgomery. Graduate Texts in Mathematics, 74. Springer-Verlag, New York, 2000.

[E] H. M. Edwards, *Riemann's zeta function*. Reprint of the 1974 original, Dover Publications, NY, 2001.

[H] Ghaith Ayesh Hiary, *Fast methods to compute the Riemann zeta Function*. arxiv.

[HB] D.R. Heath-Brown, *Private communication to A.M. Odlyzko*.

[HO] Ghaith Ayesh Hiary and A.M. Odlyzko, *The zeta function on the critical line: numerical evidence for moments, extremes, and Random Matrix Theory models*. In preparation.

[Hu] M. N. Huxley, *Area, lattice points, and exponential sums*. London Mathematical Society Monographs. New Series, 13. Oxford Science Publications. The Clarendon Press, Oxford University Press, New York, 1996.

[I] Mourad Ismail, *Classical and quantum orthogonal polynomials in one variable*. With two chapters by Walter Van Assche. With a foreword by Richard A. Askey. Encyclopedia of Mathematics and its Applications, 98. Cambridge University Press, Cambridge, 2005.

[GK] S.W. Graham and G. Kolesnik, *Van der Corputs method of exponential sums*. London Mathematical Society Lecture Note Series, 126. Cambridge University Press, Cambridge, 1991.

[Ka] Ekatherina A. Karatsuba, *Approximation of sums of oscillating summands in certain physical problems*. J. Math. Phys. 45 (2004), no. 11, 4310–4321.

[Ko] N. M. Korobov, *Exponential sums and their applications*. Translated from the 1989 Russian original by Yu. N. Shakhov. Mathematics and its Applications (Soviet Series), 80. Kluwer Academic Publishers Group, Dordrecht, 1992.

[LWY] Jianya Liu, Trevor Wooley, and Gang Yu, *The quadratic Waring-Goldbach problem*. J. Number Theory 107 (2004), no. 2, 298–321.

[M] David Mumford, *Tata lectures on theta*. With the assistance of C. Musili, M. Nori, E. Previato and M. Stillman. Progress in Mathematics, 28. Birkhuser Boston, Inc., Boston, MA, 1983.

[O] A.M. Odlyzko, *The 10^{20}-th zero of the Riemann zeta function and 175 million of its neighbors*. Manuscript. www.dtc.umn.edu/~odlyzko.

[OS] A. M. Odlyzko and A. Schönhage, *Fast algorithms for multiple evaluations of the Riemann zeta function*. Trans. Amer. Math. Soc. 309 (1988), no. 2, 797–809.

[S] A. Schönhage, *Numerik analytischer Funktionen und Komplexität*. Jahresber. Deutsch. Math.-Verein. 92 (1990), no. 1, 1–20.

[T] E.C. Titchmarsh, *The Theory of the Riemann zeta-Function*. 2nd ed., revised by D. R. Heath-Brown. Oxford University Press, Oxford, 1986.

[V] I.M. Vinogradov, *Elements of Number Theory*. Translated by S. Kravetz. Dover Publications, Inc., New York, 1954.

Department of Mathematics, University of Minnesota, 206 Church St. S.E., Minneapolis, MN, 55455. Email: hiary@math.umn.edu